



**RESEARCH DEPARTMENT**

# **A simple application to broadcasting of the results of subjective tests**

**TECHNOLOGICAL REPORT No. A-091**

UDC 519.24: 654.19

1966/43


**THE BRITISH BROADCASTING CORPORATION  
ENGINEERING DIVISION**

RESEARCH DEPARTMENT

**A SIMPLE APPLICATION TO BROADCASTING OF THE RESULTS  
OF SUBJECTIVE TESTS**

Technological Report No. A-091  
UDC 519.24: 1966/43  
654.19

R.D.A. Maurice, Dr. Ing., Ing. E.S.E., M.I.E.E.



Head of Research Department

This Report is the property of the  
British Broadcasting Corporation and  
may not be reproduced in any form  
without the written permission of the  
Corporation.

**A SIMPLE APPLICATION TO BROADCASTING OF THE RESULTS  
OF SUBJECTIVE TESTS**

Section	Title	Page
	SUMMARY . . . . .	1
1.	INTRODUCTION . . . . .	1
2.	USE OF STATISTICAL RESULTS TO FORMULATE A CONCLUSION . . . . .	2
3.	CONCLUSIONS . . . . .	3
4.	REFERENCES . . . . .	4
5.	APPENDIX . . . . .	5

## A SIMPLE APPLICATION TO BROADCASTING OF THE RESULTS OF SUBJECTIVE TESTS

### SUMMARY

*The report discusses some different scales of criteria suitable for subjective tests for broadcasting applications and mentions a method of using the statistical results of subjective tests that may make maximum use of them to aid management decision making.*

### 1. INTRODUCTION

Many occasions arise when units of objective measurement do not exist and it is necessary to rely on the opinions of human test subjects in order to determine the appropriate value of some objective parameter that may affect a feature or aspect of a broadcasting service. For example, it is necessary to be able to establish a relationship between the impairment caused to a broadcast programme by random noise and the signal-to-noise ratio measured in objective, measurable units such as decibels. Tests conducted with the object of establishing a relationship between subjective opinion and an objectively measurable quantity are usually called "subjective tests" although they are really tests involving subjective judgments. Many subjective tests use a scale of subjective units based upon a sequence of opinions that change progressively from an indication of audience satisfaction when an impairment to a broadcast programme is slight or the technical quality is good, to an indication of dissatisfaction when an impairment is significant or the technical quality is poor. Such scales of opinions or assessments are termed scales of multiple criteria. The criteria may vary in a "smooth" fashion by the use of adjectives; or they may vary in what might be termed discontinuous steps. Thus a smooth scale might be

whilst a scale with some discontinuities in it might be

TABLE 2

Criterion of technical quality of broadcast programme	Number of subjective units
perfect	1
not always perfect	2
acceptable for broadcasting	3
not always acceptable for broadcasting	4
acceptable for certain types of programme	5
unacceptable for broadcasting	6

Tables 1 and 2 use technical quality as the basis for the subjective criteria. There are other possible aspects of which impairment to technical quality is one. Thus a smooth impairment scale could be

TABLE 1

Criterion of technical quality of broadcast programme	Number of subjective units
very good	1
good	2
fairly good	3
rather poor	4
poor	5
very poor	6

TABLE 3

Criterion of technical quality of broadcast programme	Number of subjective units
impairment imperceptible	1
impairment just perceptible	2
impairment definitely perceptible	3
impairment intense	4
impairment very intense	5
impairment overwhelming	6

whilst a scale with some discontinuities in it might be

TABLE 4

Criterion of technical quality of broadcast programme	Number of subjective units
impairment imperceptible	1
impairment just perceptible	2
impairment definitely perceptible but not disturbing	3
impairment somewhat objectionable	4
impairment annoying	5
impairment unacceptable for broadcasting	6

The difficulty in constructing discontinuous scales having as many as six separate criteria can be judged from the inadequacies present in Tables 2 and 4. On the other hand, Tables 1 and 3 show a somewhat monotonous use of adjectives and moreover, they are not entirely free from discontinuities, viz. the change in wordings when going from grade 3 to grade 4.

Tables 1 to 4 show what are termed absolute scales. When two technical arrangements or systems or methods exist, it is sometimes useful to make subjective comparisons between the programmes resulting from the use of each arrangement. Thus a comparative scale has been much used by the European Broadcasting Union (EBU)<sup>1</sup>:

TABLE 5

System "A" is	Number of subjective units
much worse than	-3
worse than	-2
slightly worse than	-1
same as	0
slightly better than	1
better than	2
much better than	3
System "B"	

It should be noted that all the absolute scales, Tables 1 to 4, have an even number (six) of criteria. Odd numbered scales are also used<sup>2</sup> by many workers; perhaps the only disadvantage of them is that they have a central criterion that may tempt the test observers to use excessively when they find it difficult to make up their minds whether an impairment (for example) should be assessed to be in the favourable or the unfavourable half of the scale of criteria. It may be argued, on the other hand, that more than five criteria are difficult to define. Tables 2 and 3 may be thought to show this difficulty.

## 2. USE OF STATISTICAL RESULTS TO FORMULATE A CONCLUSION

The main object of this report is not to discuss the merits and demerits of scales of criteria, however, but to suggest a method of arriving at a useful conclusion from the statistics of any given test of technical quality or impairment to a sound or television programme. A point that is not always fully understood is that the aim of a subjective test may affect not only the type of scale of criteria used but also the method of and conclusions from the statistical analysis (more precisely, synthesis) that must invariably follow. Many tests are aimed at discovering the human reaction to some specific stimulus, such as the acuity of the eye to luminance or chromatic fine detail, or the response of the ear to tones of various pitch. Other tests, and these are of great importance to organizations that provide a public service such as broadcasting, are aimed at establishing the acceptability of some impairment whose complete elimination may be uneconomic to achieve. In the latter case, it has been the practice to conduct subjective tests, sometimes of a very extensive nature, based upon scales of multiple criteria. The criteria are each given a grade number as in Tables 1 to 4 and the statistical mean or average grade number is calculated as well as the standard deviation and, of recent years, the percentage of assessments that are found in the unsatisfactory half of the scale (grades exceeding\* the numerical value 3.5) have also been given careful consideration.

The critical question, when the figures are available for inspection, is "what degree of impairment can be regarded as acceptable for broadcasting?" No clear-cut method of arriving at an answer has yet been devised. Furthermore, a more lax standard can possibly be allowed for a new system of broadcasting than for a new impairment suddenly introduced into existing broadcast programmes. The compatibility of colour television systems is a case in point.

\* It is usual to allocate half of the number of assessments that are graded exactly 3.5 to the unsatisfactory half of the scale.

One attitude of mind that is based on very practical considerations is that which regards viewer or listener complaints as the fundamental criterion. In this connexion the scales in Tables 1 to 4 can be regarded as having an acceptable half, (grades 1 to 3) and an unsatisfactory half, (grades 4 to 6). The criterion of acceptability for broadcasting could therefore be related to the percentage of assessments classified in the unsatisfactory half of the subjective scale; that is, grades 4 to 6, or if fractions of a grade are used by some of the test observers, grades greater than 3.5.

During early BBC field trials and laboratory tests on 405-line colour television<sup>3</sup> it was suggested that if the percentage,  $P$ , of assessments of a given impairment exceeded ten, then that degree of impairment was to be regarded as unacceptable.

During the 1957/8 field trials<sup>4</sup> a first-class service area was defined as one containing no areas having more than 5 per cent of the assessments in the grade range greater than 3.5. A second-class service area was defined as one that contained no area in which the percentage of unfavourable assessments ( $>3.5$ ) exceeded 30. It can be seen that in each of these defined service areas the actual percentages of viewers who would assess their pictures as being in the grade range greater than 3.5 will be very much less than 5 and 30 respectively.

The criterion that the percentage of unfavourable opinions shall not exceed a specified figure requires a knowledge of the percentage of opinions graded greater than 3.5. Often, this percentage is directly available from the test results, but in many cases, where the tests have not been conducted by the statistical analyst, the only results available may be the statistical mean grade and the standard deviation. Again, in cases in which the size of the statistical sample of opinions (grades) is rather small, a somewhat more reliable figure for the percentage of grades exceeding 3.5 may be obtained by calculating it using the values of the mean and standard deviation rather than accepting the percentage directly from the test results. In effect, this may be regarded as a method of "smoothing" the results. In such cases Fig. 1 may be found useful. It is based upon the assumption\* that the statistical frequency distribution of the observers' grades during a subjective test follow the Normal or Gaussian law of random errors.

If the criterion  $P < 10$  per cent be taken, then Fig. 2 may be found to be useful.

\* This assumption, suggested tentatively by Monsieur L. Goussot of O.R.T.F., is shown in the Appendix to give results that are in close agreement with results from actual subjective tests.

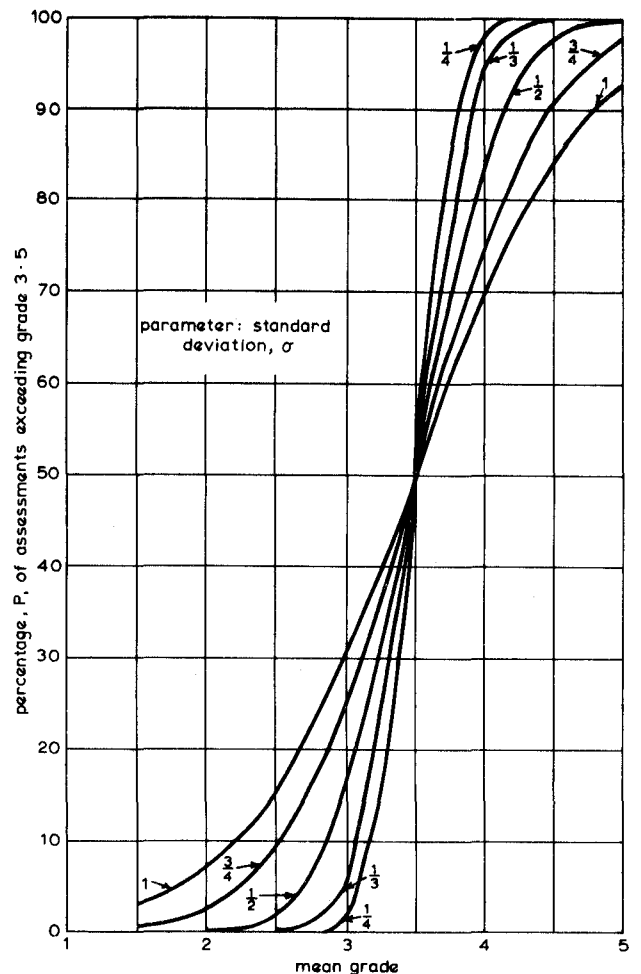


Fig. 1 - Percentage of assessments exceeding grade 3.5 as a function of mean grade and standard deviation

### 3. CONCLUSIONS

Scales of subjective criteria may be "smooth" or "discontinuous" and the types of criteria used for purely scientific investigations may differ from those used when administrative or executive decisions will be based upon the subjective test results.

The use to which the statistical results of tests may be put depends upon the type of decision (if any) that is required and for the aid of which the tests were carried out. If "suitability for broadcasting" is the ultimate criterion of an engineering system or a piece of complicated equipment to be used in the broadcasting chain, then the percentage of "lay" viewers finding the impairment caused by the equipment under test to be sufficiently strong to be classified as unsatisfactory may be used as a "go" "no-go" gauge. The choice of the actual percentage that is to act as a tolerance is, however, still very much a matter of opinion.

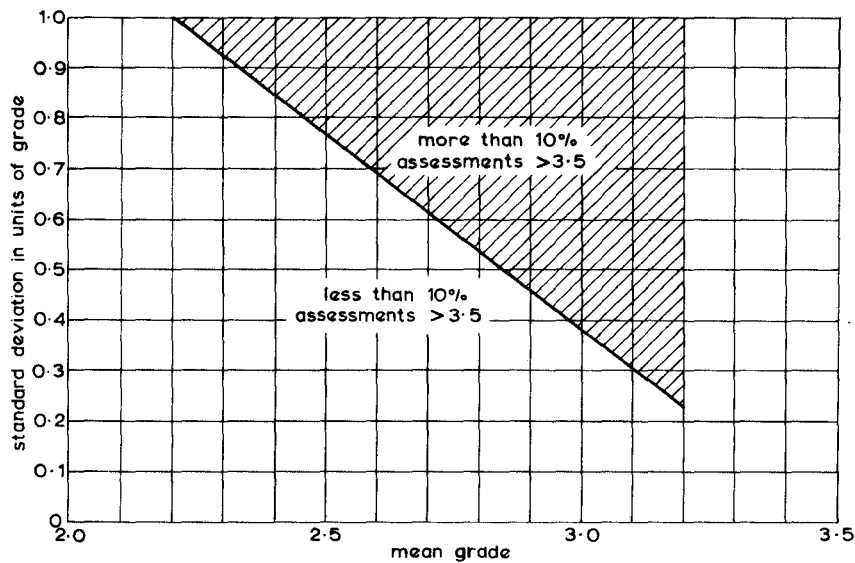


Fig. 2 - The combination of mean grade and standard deviation giving rise to percentages,  $P$ , of unfavourable opinions not exceeding 10%

#### 4. REFERENCES

1. Report of the EBU Ad-Hoc Group on Colour Television, Second Edition, February 1965.
2. PROSSER, R.D., ALLNATT, J.W., and LEWIS, N.W. 1963. Quality grading of impaired television pictures. *Proc.Instn elect.Engrs*, 1964, **III**, 3, pp. 491 - 502.
3. The BBC Colour Television Tests: an appraisal of results. BBC Engng Monogr., 1958, No. 18, Section 4, Table 3.
4. Television field trials of 405-line and 625-line systems in the UHF and VHF bands 1957-1958. London, BBC, 1960.
5. Report of the EBU Ad-Hoc Group on Colour Television, Second Edition, February 1965.



## 5. APPENDIX

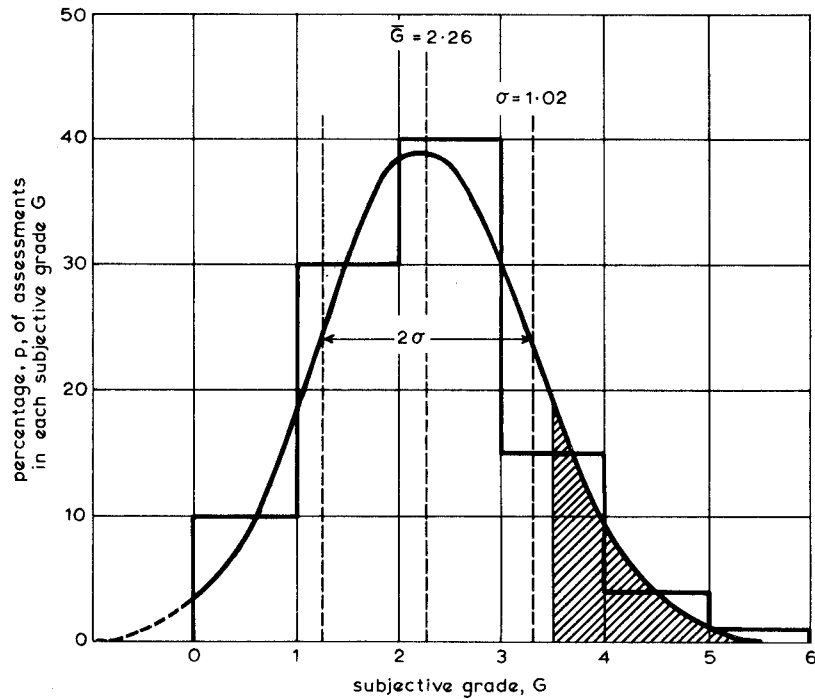


Fig. 3 - Example histogram and Gaussian curve of best-fit

In order to have a means of calculating the percentage of test observers' assessments that exceed subjective grade 3.5 when the mean grade  $\bar{G}$  (in grade units) and the standard deviation,  $\sigma$ , of a set of assessments are given, it will be assumed that the statistical frequency distribution of the assessments follows, with sufficient approximation, the Gaussian or Normal-Error law. Fig. 3, which is given only as an example, shows how the Gaussian distribution will be placed with respect to the given mean grade  $\bar{G}$  and the given standard deviation  $\sigma$ . The given information in Fig. 3 is the histogram showing the percentage  $p$  of assessments in each subjective grade. From the histogram the mean grade  $\bar{G} = 2.26$  and the standard deviation  $\sigma = 1.02$  can be calculated. The equivalent Gaussian curve thus becomes

$$p = 100/(\sigma\sqrt{2\pi}) \exp [-(G - \bar{G})^2/2\sigma^2]$$

where  $\bar{G} = 2.26$

and  $\sigma = 1.02$

In order to calculate, from a given mean value,  $\bar{G}$ , and a given standard deviation,  $\sigma$ , the percentage of assessments likely to occur in the grade range from 3.5 to 6; that is, the percentage of assessments that exceed 3.5; it is only necessary to find

the ratio of the hatched area in Fig. 3 to the total area under the Gaussian curve. If  $P$  is the desired percentage

$$P = 50(1 + \theta) \text{ for } \bar{G} \geq 3.5$$

$$P = 50(1 - \theta) \text{ for } \bar{G} \leq 3.5$$

where

$$\theta(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^x e^{-u^2} du$$

and

$$x = |\bar{G} - 3.5|/\sigma\sqrt{2}$$

$x$  and therefore  $\theta$  are both functions of  $\bar{G}$  and  $\sigma$  and, in order to facilitate the calculation of  $P$ , the curves of Fig. 1 may be used.

The question now arises, "How accurate will  $P$  be when calculated by the above process?" Some results of subjective tests that were carried out for the European Broadcasting Union programme of work in connexion with the study of colour television systems are given in a report<sup>5</sup> of that body. Table 6 shows, in each line marked "Meas." the actual value of  $P$  (column 4) taken directly from the test results. In the lines marked "Calc." the values of  $P$  calculated on the basis of the assumption of Gaussian or Normal-Error statistics are given, again in column 4.

TABLE 6

	1	2	3	4	5	6
	Viewing distance in picture heights	Mean grade $\bar{G}$	Standard Deviation	Per cent, P, of assessments > 3.5	System of colour television	Notes
Meas. Calc.	5	2.3	0.6	2 2.5	NTSC	1
Meas. Calc.	6	2.5	0.83	12 11	NTSC	2
Meas. Calc.	10	1.5	0.67	0 0	NTSC	2
Meas. Calc.	5	3	0.9	26 30	SECAM III	1
Meas. Calc.	6	4	0.93	70 70	SECAM III	2
Meas. Calc.	10	2.2	1.2	19 14	SECAM III	2
Meas. Calc.	5	3	0.8	21 26	PAL	1
Meas. Calc.	6	4.1	0.88	82 74	PAL	2
Meas. Calc.	10	2.7	1.1	26 24	PAL	2

NOTES: 1. Reference 5, page 8, lower table.  
Actual viewing distance: between 4 and 6 times picture height.

2. Reference 5, page 9, upper table.

The close agreement between the measured and calculated values of the percentages, P, of assessments in the category "greater than 3.5" seems to justify the assumption of Gaussian statistics, at least for the calculation of the quantity P.